重点科普下中心极限定理。

> **1.**什么是中心极限定理

有时候统计概率就像魔术一样，能够从少量数据中得出不可思议的强大结论。我们只需要对1000个美国人进行电话调查，就能去预测美国总统大选的得票数。

通过对为肯德基提供鸡肉的加工厂生产的100块鸡肉进行病毒（沙门氏菌）检测，就能得出这家工厂的所有肉类产品是否安全的结论。

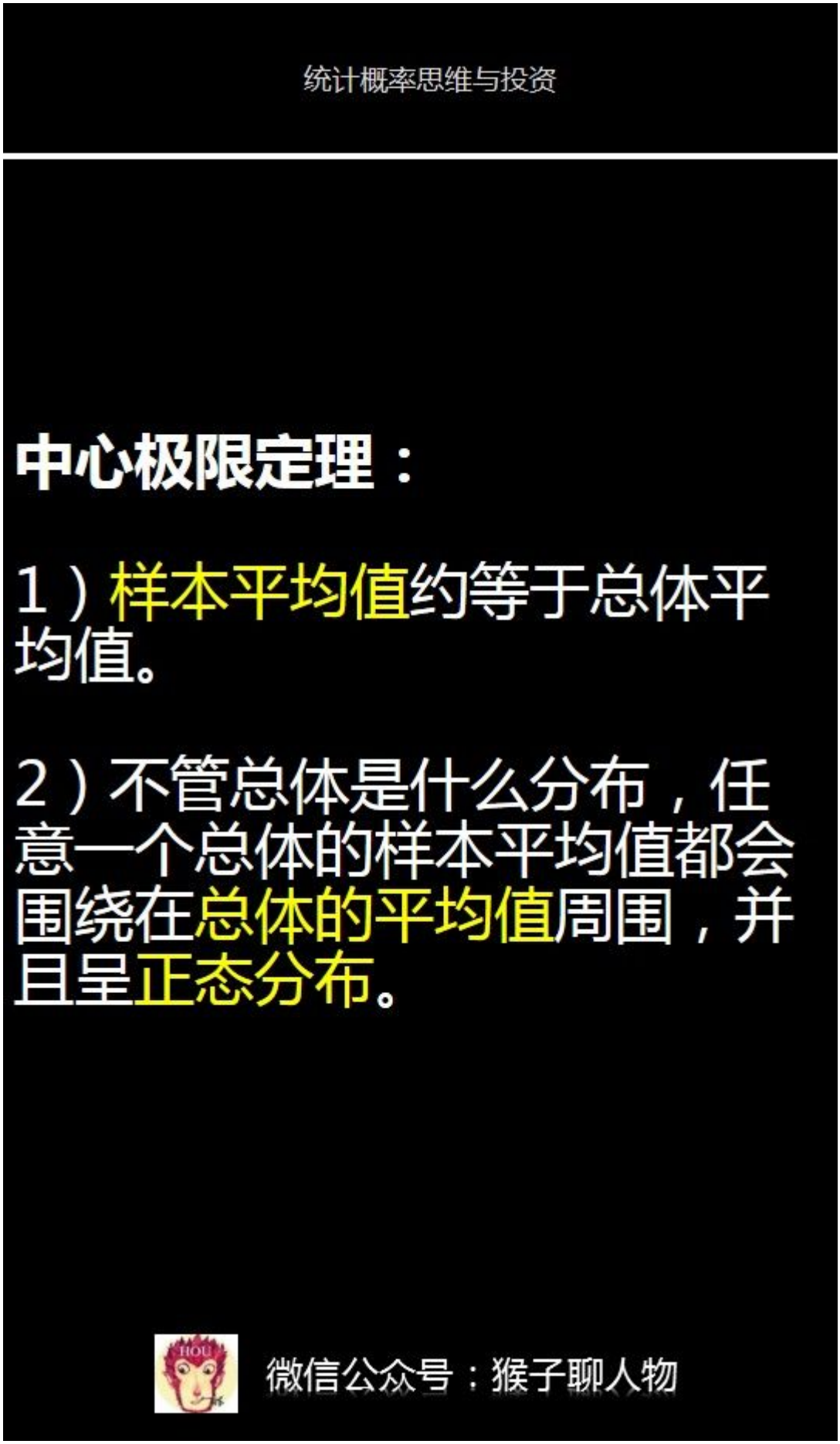
这些“一概而论”的强大能力，到底是从哪里来的？

这背后的秘密武器就是统计概率的第2大护法：中心极限定理。第1大护法我在第3讲《投资赚钱与概率》中有讲过就是：大数定律。

中心极限定理是许多统计活动的“动力源泉”，这些活动存在着一个共同的特点，那就是使用样本对总体进行估计，例如我们经常看到的民意调查就是这方面的经典案例。

那么，什么是中心极限定理呢？

> 中心极限定理是说： 样本的平均值约等于总体的平均值。 不管总体是什么分布，任意一个总体的样本平均值都会围绕在总体的整体平均值周围，并且呈正态分布。



现在看了这2句话，你肯定会说：猴子，请说人话。

别担心，我将拆开这2句话来慢慢为你聊清楚什么是中心极限定理。

假设有一个群体，如我们之前提到的清华毕业的人，我们对这类人群的收入感兴趣。怎么知道这群人的收入呢？我会做这样4步：

第1步.随机抽取1个样本，求该样本的平均值。例如我们抽取了100名毕业于清华的人，然后对这些人的收入求平均值。

该样里的100名清华的人，这里的100就是该样本的大小。

有一个经验是，样本大小必须达到30，中心极限定理才能保证成立。

第2步.我将第1步样本抽取的工作重复再三，不断地从毕业的人中随机抽取100个人，例如我抽取了5个样本，并计算出每个样本的平均值，那么5个样本，就会有5个平均值。

这里的5个样本，就是指样本数量是5。

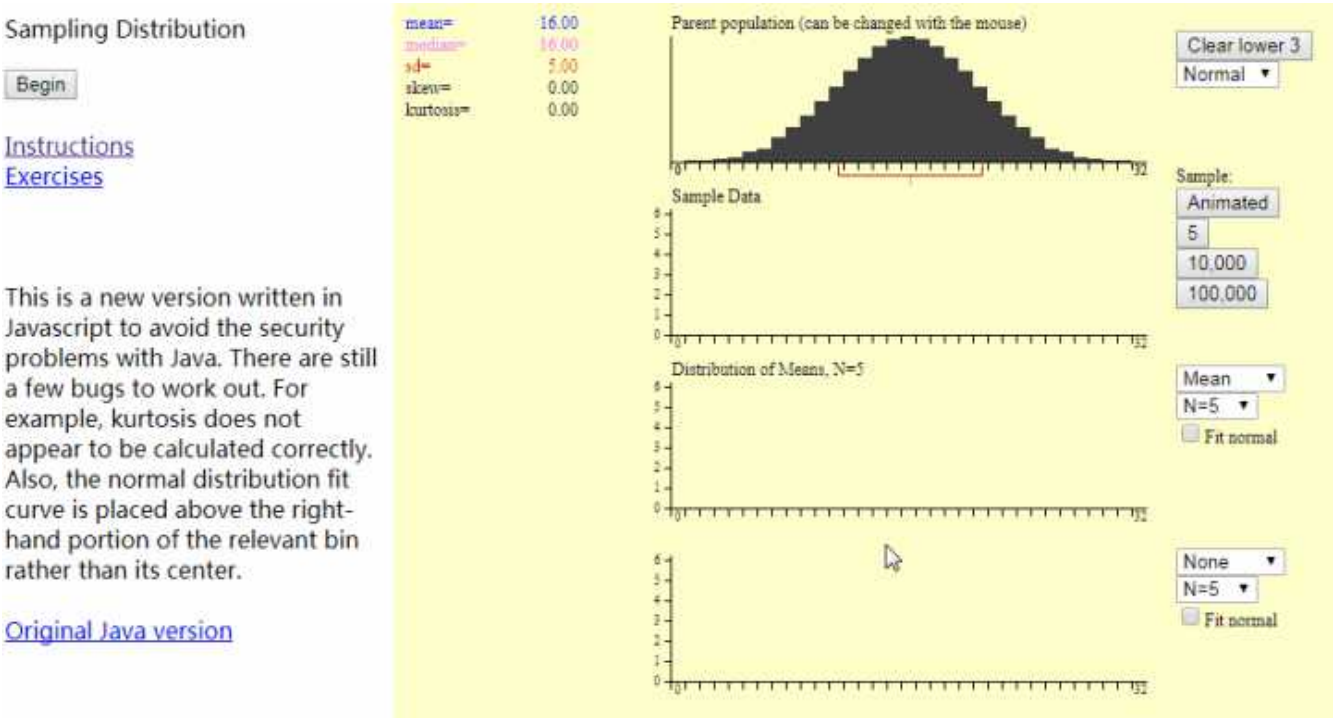
第3步.根据中心极限定理，这些样本平均值中的绝大部分都极为接近总体的平均收入。有一些会稍高一点，有一些会稍低一点，只有极少数的样本平均值大大高于或低于群体平均值。

第4步.中心极限定理告诉我们，不论所研究的群体是怎样分布的，这些样本平均值会在总体平均值周围呈现一个正态分布。

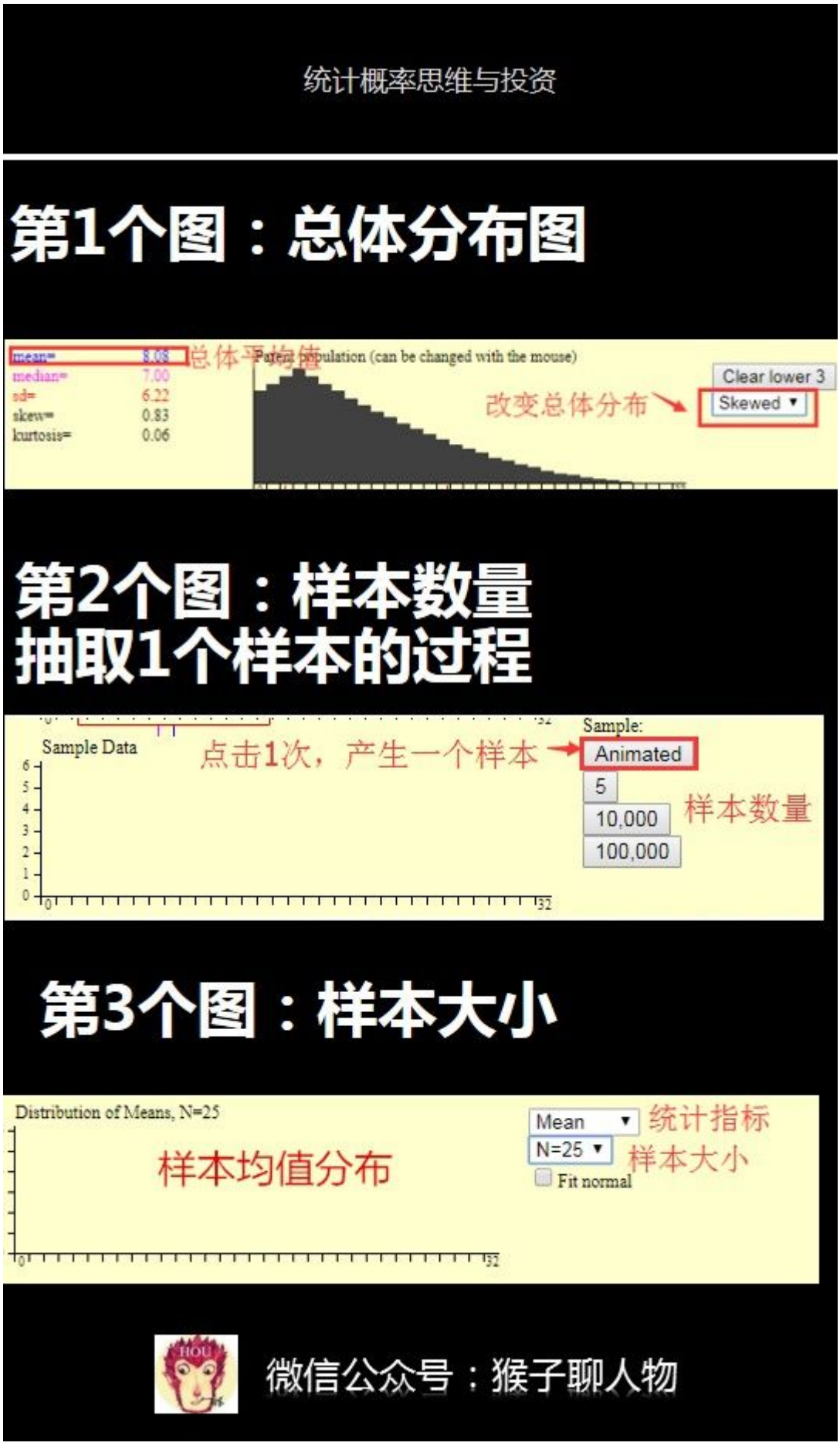
现在，我将介绍一个小程序来演示中心极限定理，通过多种方式，我们来熟悉这一重要知识。

这个小程序的演示地址我已经放到这次课程的学习道具中。下面我会介绍下这个演示程序，方便你最后可以自己动手亲自操作。

（演示中心极限定理：http://onlinestatbook.com/stat_sim/sampling_dist/index.html）



打开这个程序以后，你发下面下面图片中的三个统计图。



最上面的图是总体分布图，左边是一些统计指标，这里我们只要关注总体平均值就可以了，通过选择第1个图中红色箭头表示的地方来改变总体的分布，你可以选择总体是正态分布，或者非正态分布。

第2个统计图用来模拟产生一个样本的过程。每点击1次红色箭头标识的地方，就生成一个样本。这样通过重复点击这个按钮，你可以生成多个样本。

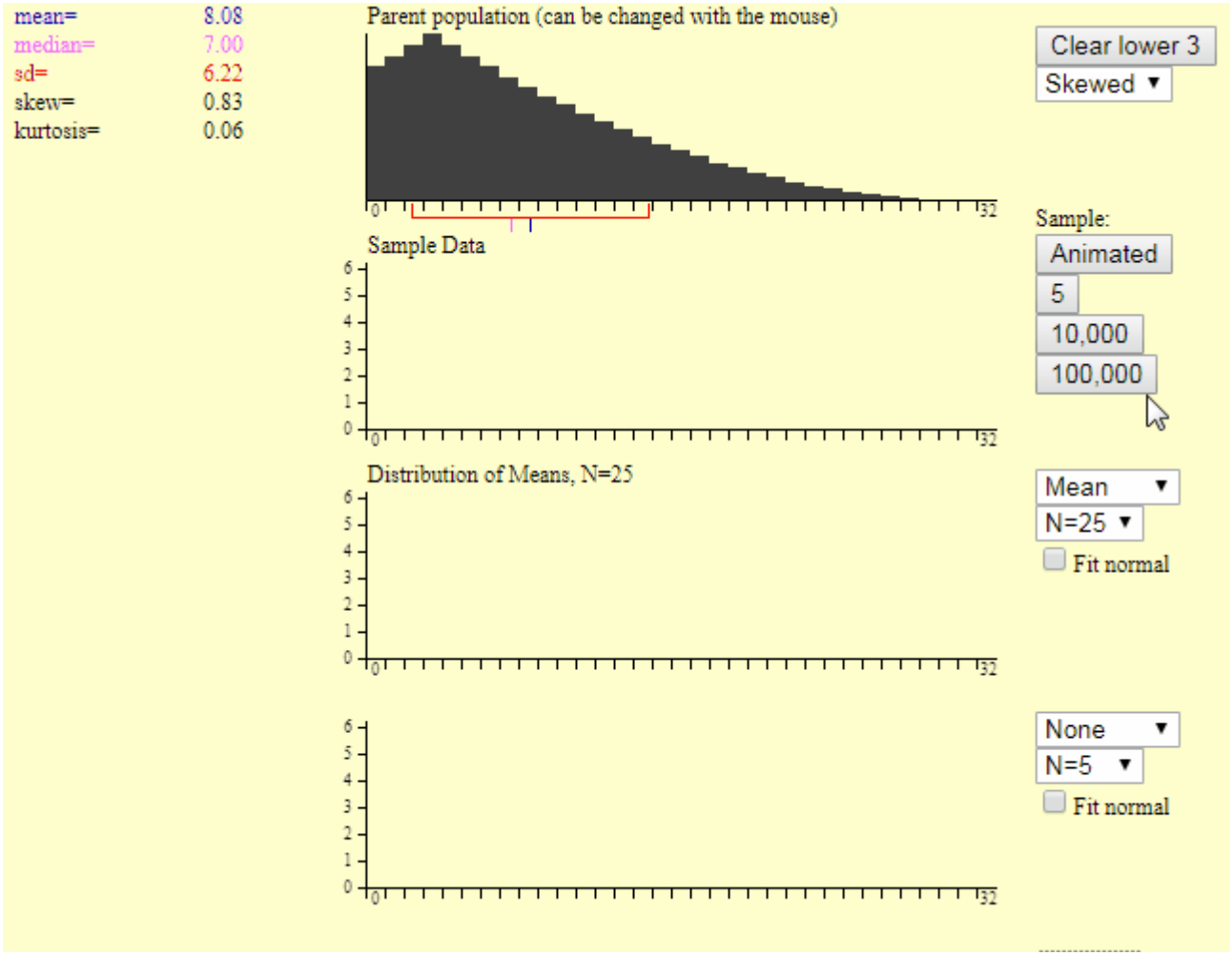
这个按钮下面的5,1千，1万数字表示，你点击该按钮，一次性帮你生成的样本数量。

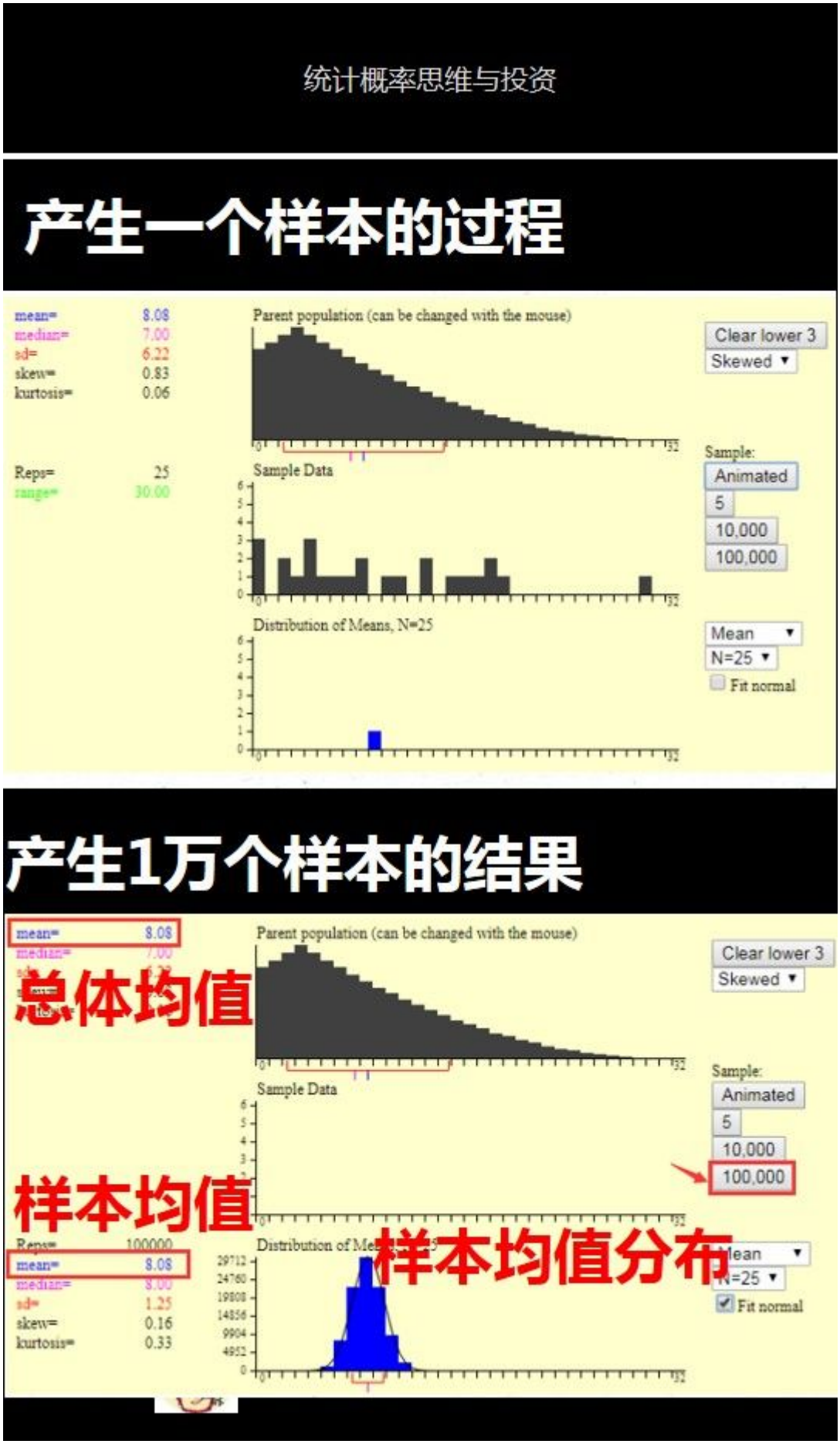
为了看清楚每个样本产生的过程，建议一开始通过点击第2个图中红色箭头那里的按钮来自己生成多个样本。

第三个图是，样本均值分布图。左边第一个你可以选择统计指标，这里我们选择平均值就可以。

第2个N是表示样本的大小，即一个样本里面有多少个数据。这里可供我们选择的最大值是25。

现在我点击第2个图中红色箭头处的按钮，便产生了下面图片的样本均值图





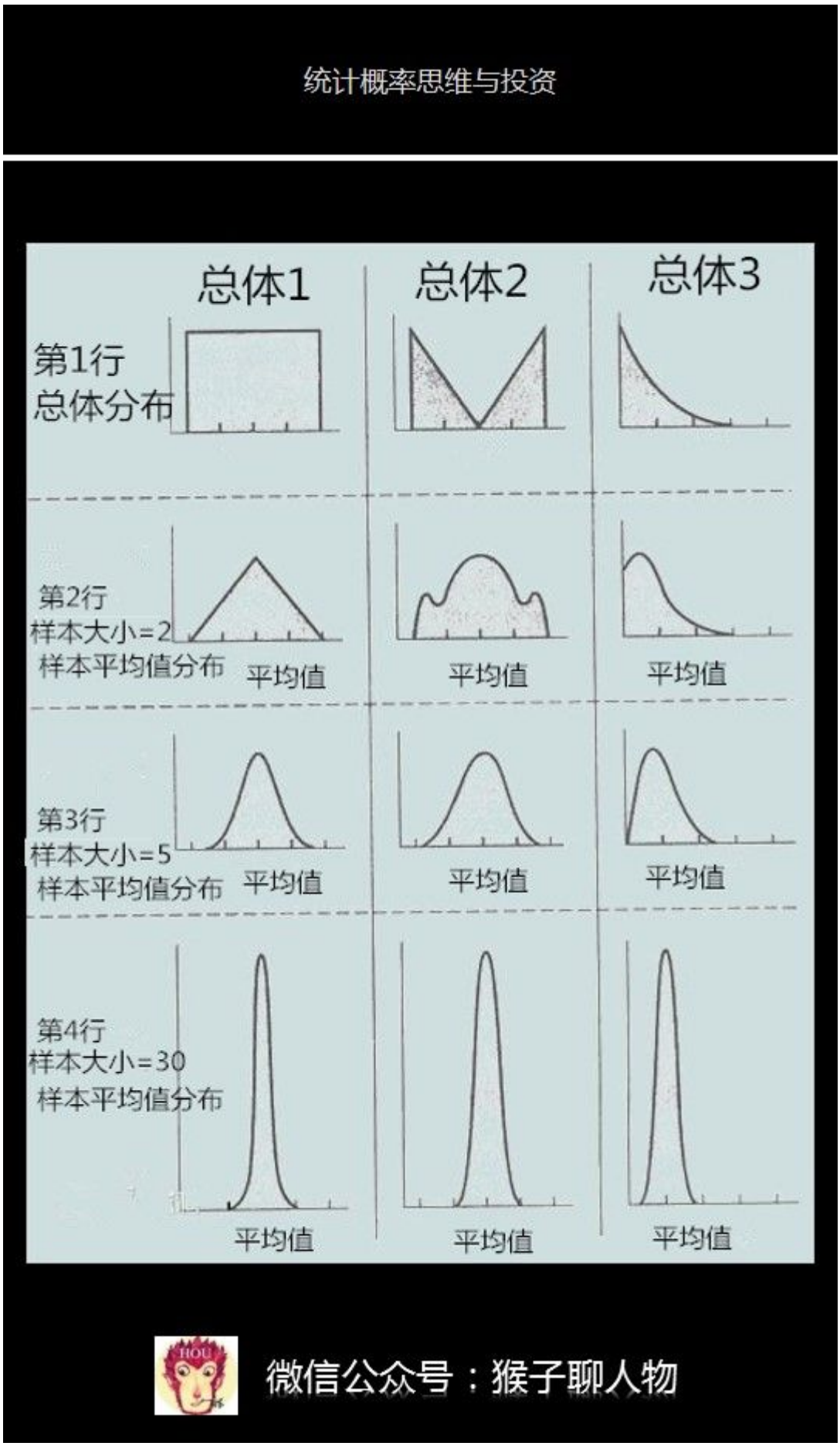
这里的第1个图是产生一个样本的过程，第2个图是产生1万个样本的结果。我们可以发现：

1）样本平均值约等于总体平均值。

2）不管总体是什么分布，任意一个样本平均值都会围绕在总体平均值周围，并且呈正态分布。

这就是中心极限定理，就是这么2句话。

下面图片也可以完美的解释中心极限定理。



这里第1行是3种不同分布类型的总体，用于比较不同类型下的样本平均值分布。

第2行每个样本大小是2，然后对每个样本求平均值，横轴表示每个样本的平均值，纵轴表示该平均值出现了多少次，最后平均值分布很不规则

第3行每个样本大小是5，然后对每个样本求平均值，最后平均值分布有点接近于正态分布，但是总体3对应的第3行却不是正态分布。

第4行每个样本大小是30，然后对每个样本求平均值，最后平均值分布是正态分布。

这也验证了中心极限定律，不管总体是什么分布，任意一个总体的样本平均值都会围绕在总体的平均值周围，并且呈正态分布。

现在你已经知道了中心极限定理的大体意思，下面图片我们通过几个案例来实践应用下。

> **2.** 中心极限定理应用案例

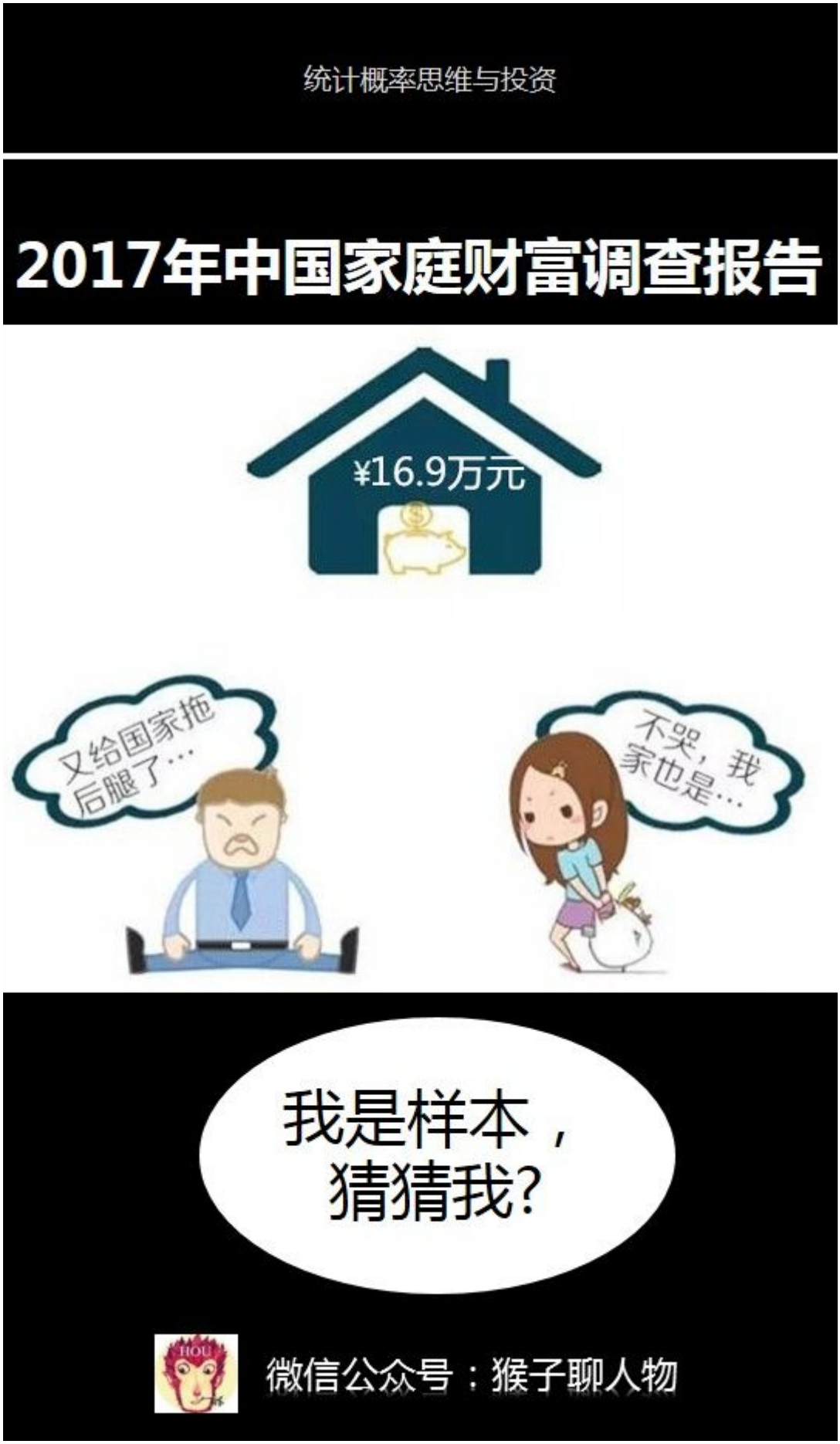
根据《2017年中国家庭财富调查报告》调查数据显示，2016年我国家庭人均财富大约为16.9万元（169077元）

（其中，房产净值是家庭财富最重要的组成部分。在全国家庭的人均财富中，房产净值的占比为65.99%）

现在假设我们随机抽样1000个中国家庭并询问他们的年收入。

根据已知的这些信息，从中心极限定理出发，你能得出什么信息？

下面我们一起来用中心极限定理进行推理。



1）根据中心极限定理，我们可以得出的第**1**个结论是：用样本来估计总体。

任何一个样本的平均值将会约等于其所在总体的平均值。

例如你久居大城市，过年回老家，大街上遇到了邻居大妈，虽然20年没见你，邻居大妈还是一眼认出你了，这不是隔壁老王家的孩子嘛，长的真带劲。

这里，你爸妈就是总体，你就是你爸妈的样本，和你爸妈长的相似。

同样的，一个正确抽取的家庭样本应该能够反映中国所有家庭的情况，里面会包含收入高的公司高管，也会包括普通的员工，快递小哥、警察以及其他人，这些人出现的频率与他们在人口构成中的占比相关。

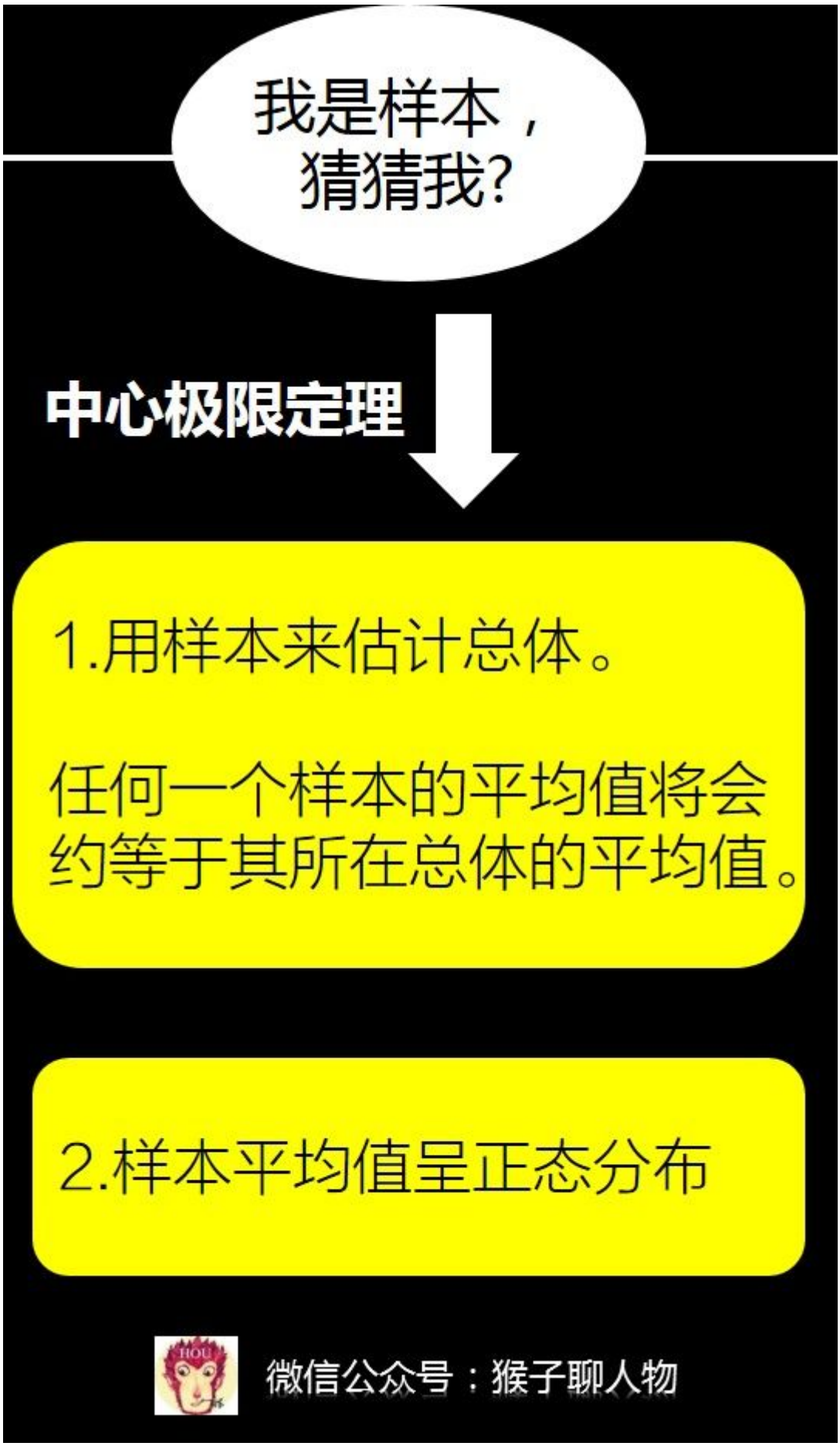
因此，我们能够推测，这个包含1000个中国家庭代表性样本的家庭财富的平均值约等于总体的平均值。

2） 样本平均值呈正态分布

在这个例子中，样本平均值将会围绕着群体平均值（也就是16.9万元）形成一条正态分布曲线。记住，群体本身的分布形态并不重要，中国家庭收入的分布曲线并非正态分布，但样本平均值的分布曲线却是正态分布。

如果我们连续抽取100次包含1000个家庭的样本，并将它们的平均值的出现频率在坐标轴上标出，那么我们基本可以确定在总体平均值周围将会呈现正态分布。

取样次数越多，结果就越接近正态分布；而且样本大小越大，分布就越接近正态分布。


我是样本，
猜猜我?
中心极限定理
1.用样本来估计总体。
任何一个样本的平均值将会
约等于其所在总体的平均值。
2.样本平均值呈正态分布
HOU
微信公众号：猴子聊人物

> **3.** 如何用样本估计总体？

现在我们已经可以用样本来估计出总体平均值。现在我想用样本来估计出总体的标准差，该怎么办呢？

我们已经知道，一个数据集的标准差是数值与平均值的偏离程度。

当你选择一个样本后，相比总体，你拥有数据的数量是变少了，因此，与总体中的数值偏离平均值的程度相比，样本中很有可能把较为极端的数值排除在外，这样使得数值更有可能以更紧密的方式聚集在均值周围。

也就是说，样本的标准差要小于总体标准差。

所以，为了更好的用样本估计总体的标准差，统计学家就将标准差的公式做了像下面图中公式中这样的改造。

某个数据集的标准差σ

$$= \sqrt{\frac{\sum (x-\mu)^2}{n}}$$

样本标准差
（用样本估计总体标准差）

$$s = \sqrt{\frac{\sum (x-\mu)^2}{n-1}}$$

即原来的标准差公式是除以n，为了用样本估计总体标准差，现在是除以n-1。这样就是的标准略大。一般用字幕s表示用样本估计出的总体标准差。

很多书上都会把除以n-1的标准差叫做样本标准，其实会给很多人造成误解。其实这个样本标准差的目的是用于估计总体标准差。

你可能会疑惑，那我什么时候标准差除以n还是n-1呢？

那就要看你使用标准差的目的是什么。

如果你只是想计算一个数据集的标准差，那么就除以n，例如你有100个毕业与清华人的收入，只是想了解这100个人构成的数据集的波动大小，那你就用除以n的标准差公式。

如果你想把这100个人当成一个样本，用这个样本来估计出总体（所有毕业与清华人的收入）的标准差，那么就除以n-1的标准差公式。

我们在看下什么是标准误差？

标准差是用来衡量数据集的波动大小。比如毕业于清华大学所有人的收入分布。

标准误差其实也是标准差，只不过它是所有样本平均值的标准差。

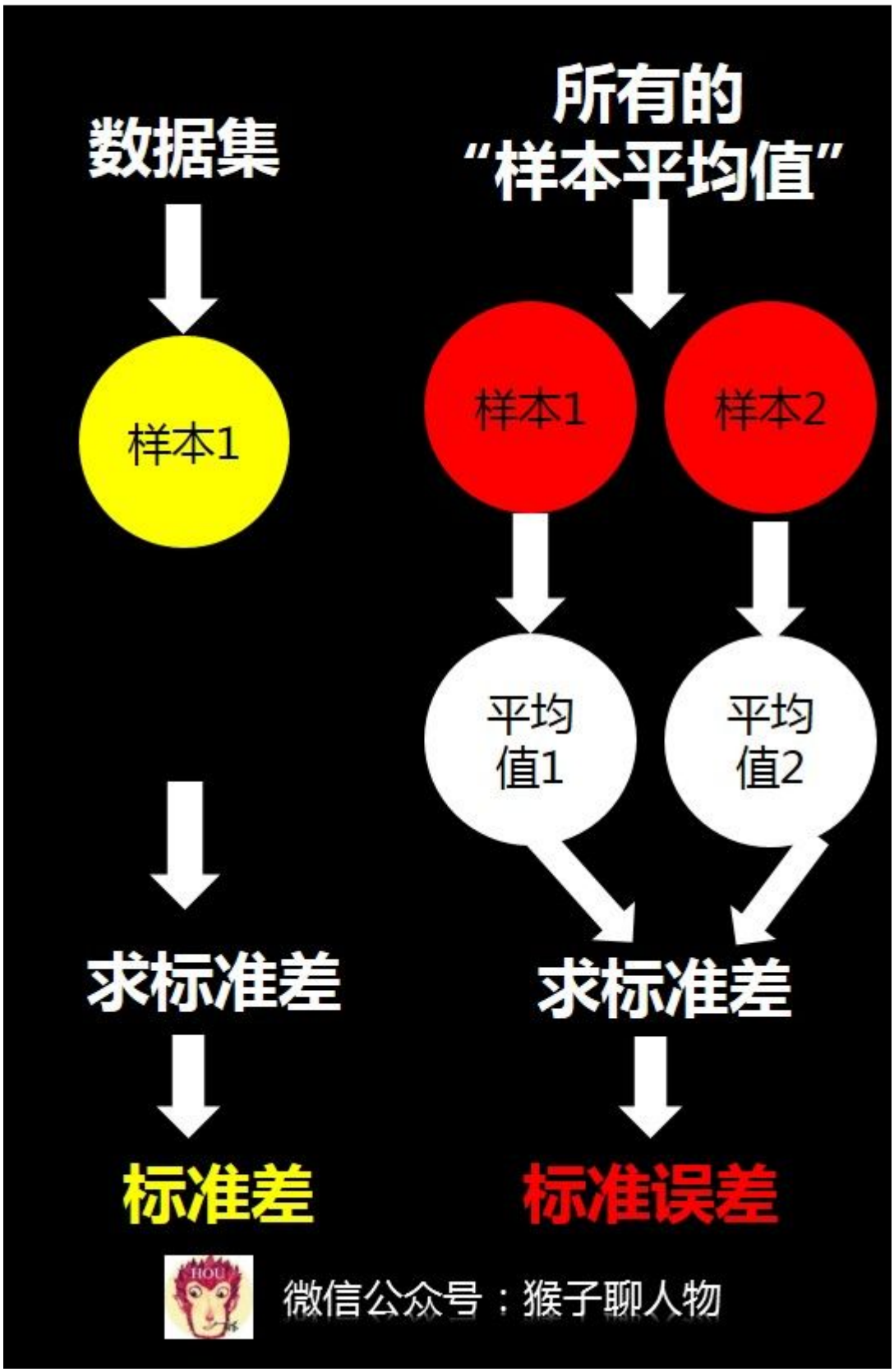
结合我刚才给的图片中的例子就更容易理解了。

如果我从毕业于清华大学中抽取100个人作为样本1，然后我计算出标准差。那么这个标准差就是用来描述这100个人组成的数据集的波动大小。

我连续刚才重复抽取样本的动作，最后抽取出2个样本，每个样本都有100个人。对每个样本计算平均值，这样就有2个平均值。

这2个平均值其实组成了1个新的数据集，就是所有的“样本平均值”。然后对这2个平均值数据计算出标准差。就是标准误差。

统计概率思维与投资


数据集
样本1
求标准差
标准差
所有的
“样本平均值”
样本1
样本2
平均
值1
平均
值2
求标准差
标准误差
微信公众号：猴子聊人物

你看，标准误差其实也是标准差，只不过它的计算对象是所有的“样本平均值”。所以，标准误差是用来衡量样本平均值的波动大小。

其实，计算标准误差有个简单的公式。下面图片我们一起看下。

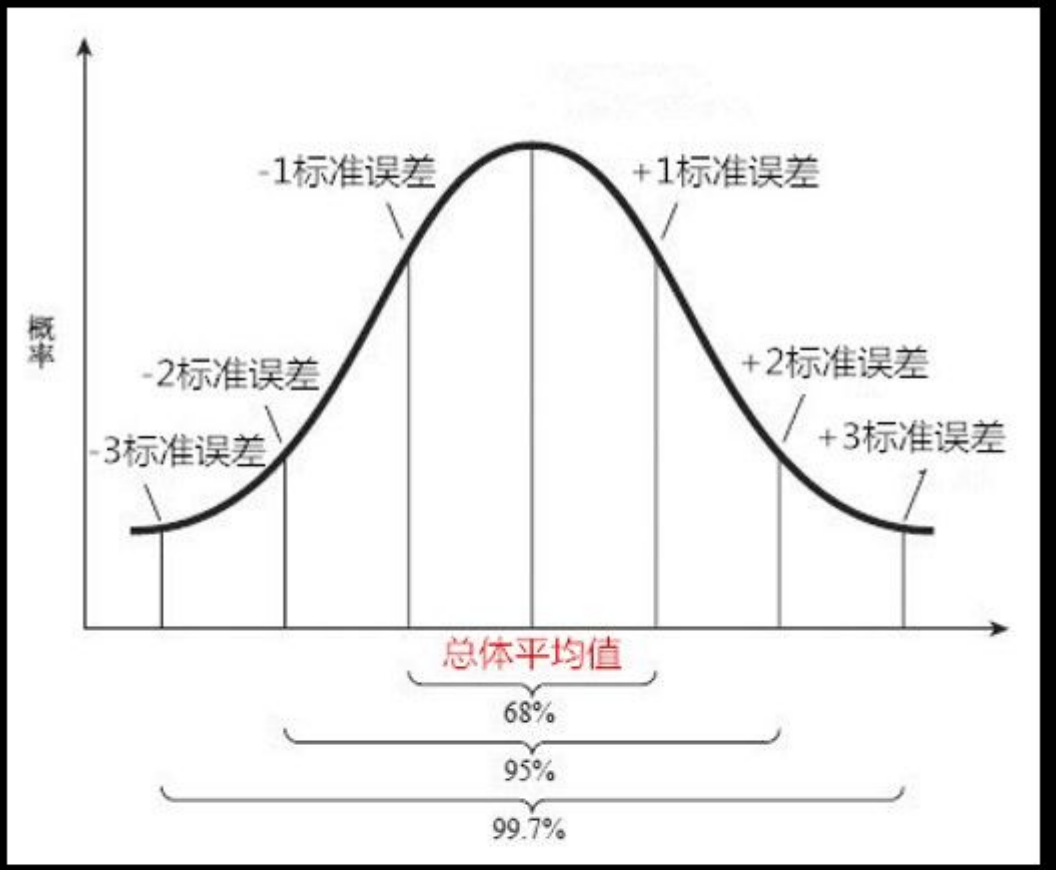
统计概率思维与投资

标准误差

$$SE = \frac{s}{\sqrt{n}}$$

s 总体标准差

n 样本大小



样本平均值概率图

微信公众号：猴子聊人物

标准误差SE等于总体标准差除以n的开方。但是我们不知道总体标准差怎么办。其实前面我们已经讲了可以用样本来估计出总体标准差的公式s。

根据中心极限定理，我们知道样本平均值是呈正态分布的，那么我们便可以通过这里图片中的样本平均值概率图来获得推理所需的“超能力”。

看到这个图是不是很熟悉，这个图其实就是前面我们讲过的正态分布概率图，只不过这里的横轴是样本平均值的大小，纵轴是该平均值出现的概率。这里是标准误差。

在前面介绍正态分布的时候，我们已经知道了正态分布的一个奇特超能力，应用到样本正态分布上，那就是：

1）有68%的样本平均值会在总体平均值一个标准误差的范围之内

数值范围（总体平均值-1个标准误差，总体平均值+1个标准误差）

2）有95%的样本平均值会在总体平均值的两个标准误差的范围之内

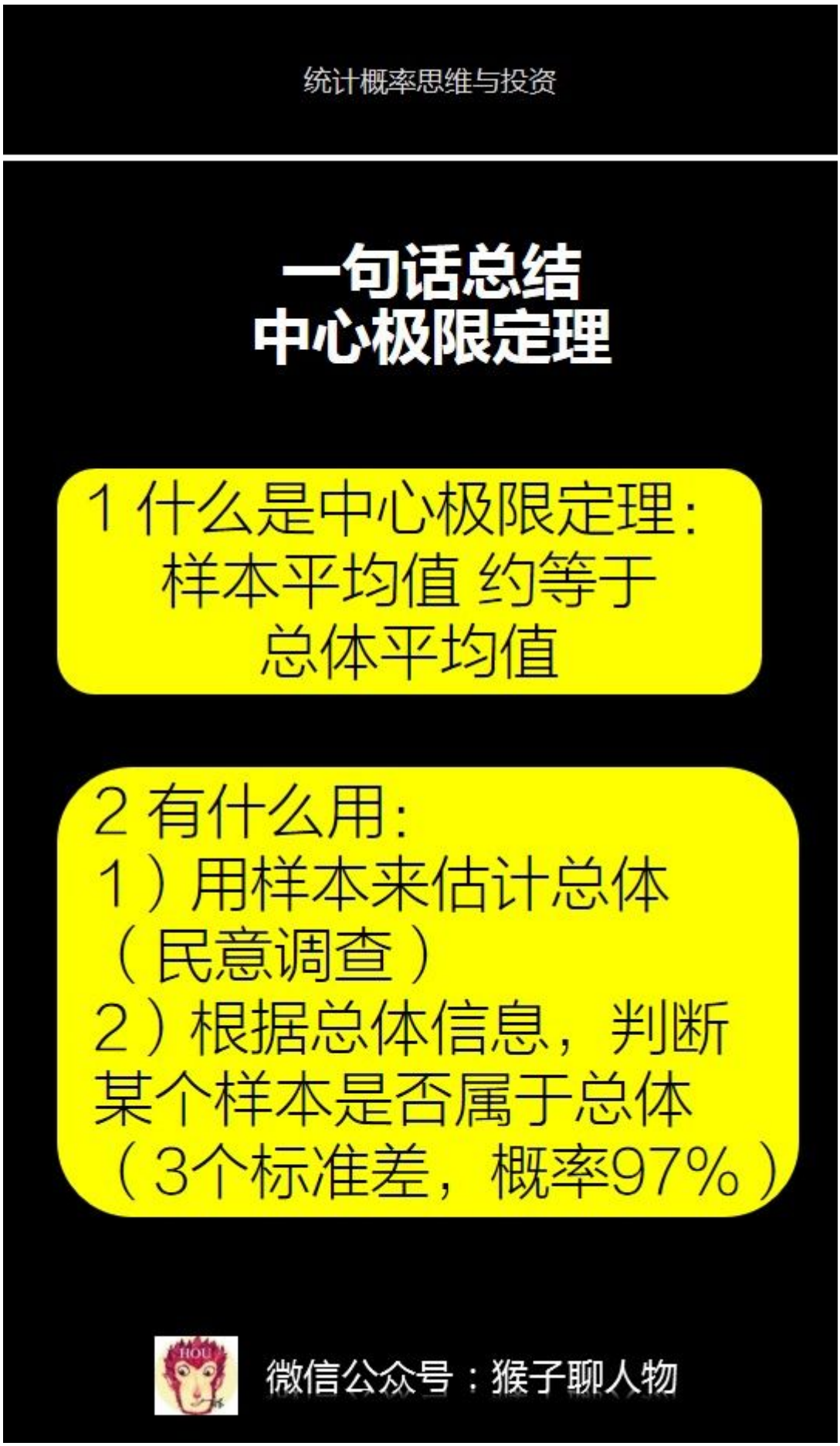
（总体平均值-2个标准误差，总体平均值+2个标准误差）

3）有99.7%的样本平均值会在总体平均值3个标准误差的范围之内。

（总体平均值-3个标准误差，总体平均值+3个标准误差）

假如某个样本的平均值减去总体的平均值，大于3个标准误差。根据99.7%的样本平均值会处于总体平均值3个标准误差的范围内，因此我们可以得出该样本不属于总体。

> **4.**一句话总结中心极限定理



中心极限定理也就是这么两句话：

1）任何一个样本的平均值将会约等于其所在总体的平均值。

2）不管总体是什么分布，任意一个总体的样本平均值都会围绕在总体的平均值周围，并且呈正态分布。

2.中心极限定理有什么用呢？

1）在没有办法得到总体全部数据的情况下，我们可以用样本来估计总体

如果我们掌握了某个正确抽取样本的平均值和标准差，就能对估计出总体的平均值和标准差。

举个例子，如果你是北京西城区的领导，想要对西城区里的各个学校进行教学质量考核。

同时，你并不相信各个学校的的统考成绩，因此就有必要对每所学校进行抽样测试，也就是随机抽取100名学生参加一场类似统考的测验。

作为主管教育的领导，你觉得仅参考100名学生的成绩就对整所学校的教学质量做出判断是可行的吗？

答案是可行的。中心极限定理告诉我们，一个正确抽取的样本不会与其所代表的群体产生较大差异。也就是说，样本结果（随机抽取的100名学生的考试成绩）能够很好地体现整个群体的情况（某所学校全体学生的测试表现）。

当然，这也是民意测验的运行机制所在。通过一套完善的样本抽取方案所选取的1200名美国人能够在很大程度上告诉我们整个国家的人民此刻正在想什么。

2）根据总体的平均值和标准差，判断某个样本是否属于总体

如果我们掌握了某个总体的具体信息，以及某个样本的数据，就能推理出该样本是否就是该群体的样本之一。

通过中心极限定理的正态分布，我们就能计算出某个样本属于总体的概率是多少。如果概率非常低，那么我们就能自信满满地说该样本不属于该群体。

这也是统计概率中假设检验的原理，假设检验我会在之后的课程中具体介绍。

关于大数定律可以看这里：猴子：大数定律具体是个什么概念？